Ciência da Computação – Universidade Federal do Agreste de Pernambuco (BCC-UFAPE) – Garanhuns – PE – Brasil

Projeto da Disciplina

Disciplina Optativa: Fundamentos de Ciência de Dados

Professores Responsáveis
Ryan Azevedo & Assuero Ximenes
{ryan.azevedo, assuero.ximenes}@ufape.edu.br

Parte 3 – Estatística e Aprendizado de Máquina

Todo o projeto deve ser implementado/desenvolvido em Python. A parte de visualização de Dados deve ser em Python ou em Python e *Java Script*, os gráficos podem ser gráficos dinâmicos ou estáticos. Podem usar funções do Python para implementar as estatísticas e os algoritmos necessários.

Engenharia de Dados (Para Todo o Projeto)

A primeira etapa do projeto consiste na Engenharia de Dados. As atividades que devem ser realizadas nessa etapa são:

- Extração dos dados dos jogos e jogadores (Apenas do seu time e dos seus Jogadores) da NBA temporada 23-24 (passada) e 24-25 (atual).
- Nas bases [isso pode ser e [deve ser] feito durante o ciclo de desenvolvimento] verificar dados ausentes, valores redundantes, ruidosos, inconsistentes, enviesados e valores outliers caso existam.
- Transformar os dados, verificar tipos, converter dados qualitativos para quantitativos, excluir ou não utilizar colunas do *Dataset* que não impactam [ainda] no projeto [Ex.: salário do jogador], normalizar valores numéricos quando necessário, preencher dados ausentes com dados reais e disponíveis pela entidade/empresa/corporação NBA [no caso de placar de jogos não extraídos automaticamente].
- Como *amostras de dados* vamos utilizar todos os dados do *Dataset* (100% dos dados que nos interessa). [Não estamos preocupados com desempenho computacional].
- O procedimento de separação [Treinamento e teste] do conjunto de dados será realizado apenas durante o desenvolvimento dos modelos preditivos e estatísticos que usaremos. Aqui usaremos técnicas automáticas para evitar o *data leakage*.

Com os dados coletados, passamos para a fase de desenvolvimento das funcionalidades relacionadas ao time/equipe.

Modelos Estatísticos - Funcionalidades/Requisitos

Aplique nos dados dos jogadores [dados extraídos e tratados nas partes 1 e 2 desse projeto] os modelos estatísticos e modelos preditivos pedidos.

- **Método de Gumbel:** modela eventos extremos.

  RF1 – Precisamos modelar e prever eventos extremos, assim precisamos verificar em cima dos dados que possuímos as probabilidades de ocorrência de pontuação, assistências e rebotes máximos e mínimos. Como *pergunta guia* responda:
  - Probabilidade de marcar acima de X [pontos, rebotes, assistências]?
  - Probabilidade de atingir ou exceder X [pontos, rebotes, assistências]:
  - Probabilidade de atingir ou ficar abaixo de X [pontos, rebotes, assistências]?
  - Proporção de valores menores ou iguais a X [pontos, rebotes, assistências]:
  - Valores menores que X
  - Proporção de valores menores que X:

  RF2 – Apresente gráficos que facilitem a visualização dos extremos e das respostas as perguntas realizadas no RF1. Use gráficos do seu interesse.

- **Regressão Linear:** variáveis dependentes e independentes

  RF3 – Possível uso de variáveis independentes: tempo que o jogador passou em quadra, arremessos tentados e turnovers. Variáveis dependentes, pontos, assistências e rebotes. Divida os dados de teste e treinamento.
  - Responda;
    - As probabilidades de o jogador marcar acima e abaixo da média, mediana, moda, máximo e mínimo para pontos, rebotes e assistências.

  RF4 – Apresente gráficos que facilitem a interpretação das previsões como matriz de confusão, gráficos de probabilidade predita, curva roc, gráficos de coeficientes, etc.

  .

- **Regressão Logística:** variável alvo

  RF5 – Possível uso de variáveis independentes: tempo que o jogador passou em quadra, arremessos tentados e turnovers. Variáveis dependentes, pontos, assistências e rebotes. Divida os dados de teste e treinamento.

  RF6 – Apresente gráficos que facilitem a interpretação das previsões, como matriz de confusão, gráficos de probabilidade predita, curva roc, gráficos de coeficientes, etc.

- **GAMLSS:** *Generalized Additive Models for Location Scale and Shape*

RF7 – Prever uma quantidade X [pontos, rebotes, assistências] no próximo jogo. Use a função PoissonGAM e a função LinearGAM da biblioteca pygam.

- Responda:
  - As probabilidades de o jogador marcar acima e abaixo da média, mediana, moda, máximo e mínimo para pontos, rebotes e assistências, além de predizer **exatamente** um valor x.

RF8 – Apresente gráficos que facilitem a interpretação das previsões, como matriz de confusão, gráficos de probabilidade predita, curva roc, gráficos de coeficientes, etc.

OBS.: Para segunda etapa do projeto final, faremos uso de aprendizado profundo de máquina e implementaremos a Rede Neural LSTM. Não é preciso implementar até a entrega dessa etapa do projeto (Primeira avaliação), caso o aluno deseje, pode iniciar essa implementação de forma livre e sem requisitos impostos.